University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2007 *Paper* 17

Statistical Methods for Inference of Genetic Networks and Regulatory Modules

Hongzhe Li*

*hongzhe@mail.med.upenn.edu

Statistical Methods for Inference of Genetic Networks and Regulatory Modules

Hongzhe Li

Abstract

Large-scale microarray gene expression data, motif data derived from promotor sequences, genome-wide chromatin immunoprecipitation (ChIP-chip) data, DNA polymorphism data and epigenomic data provide the possibility of constructing genetic networks or biological pathways, especially regulatory networks. In this paper, we review some new statistical methods for inference of genetic networks and regulatory modules, including a threshold gradient descent procedure for inference of Gaussian graphical models, a sparse regression mixture modeling approach for inference of regulatory modules, and the varying coefficient model for identifying regulatory subnetworks by integrating microarray time-course gene expression data and motif or ChIP-chip data. We present the statistical formulations of the problems, statistical methods, and results from analysis of real data sets. Areas of future research are also discussed.

Statistical Methods for Inference of Genetic Networks and Regulatory Modules

A Chapter in "Analysis of Microarray Data" edited by Dehmer and Emmert-Streib, to be published by Wiley-VCH.

Hongzhe Li
Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6021, U.S.A.
email: hongzhe@mail.med.upenn.edu

Abstract

Large-scale microarray gene expression data, motif data derived from promotor sequences, genome-wide chromatin immunoprecipitation (ChIP-chip) data, DNA polymorphism data and epigenomic data provide the possibility of constructing genetic networks or biological pathways, especially regulatory networks. In this paper, we review some new statistical methods for inference of genetic networks and regulatory modules, including a threshold gradient descent procedure for inference of Gaussian graphical models, a sparse regression mixture modeling approach for inference of regulatory modules, and the varying coefficient model for identifying regulatory subnetworks by integrating microarray time-course gene expression data and motif or ChIP-chip data. We present the statistical formulations of the problems, statistical methods, and results from analysis of real data sets. Areas of future research are also discussed.

1 Introduction

The completion of the human genome project and the development of many high-throughput genomic technologies make it possible to systematically define the organization and function of gene, protein and metabolite networks. Large-scale microarray gene expression data, promotor sequences data and genome-wide chromatin immunoprecipitation (ChIP-chip) data provide the possibility of learning gene regulation and constructing the gene regulatory networks and pathways or cellular networks (Ideker *et al.*, 2001; Friedman, 2004; Das *et al.*, 2006). In a recent review, Bansal *et al.* (2007) summarized the methods for inferring genetic networks into two broad classes: those based in the "physical interaction" approach that aims at identifying interactions among transcription factors and their target genes and those based on the "influence interaction" approach that aims to relate the expression of a gene to the expression of the other genes in the cell, rather than relating it to the sequence motif found in its promotor. In this paper, we review some recently developed statistical methods for several problems related to inferences of genetic network and regulatory modules, including both "physical interaction" networks using diverse data sets and "influence interaction" networks using gene expression data alone.

Early research on gene expression analysis has mainly focused on using clustering analysis to identify co-regulated genes (Tavazoie *et al.*, 1999). Recently, some efforts have been devoted to developing probabilistic models for modeling regulatory and cellular networks based

on genome-wide high-throughout data, including both Bayesian network modeling (Friedman, 2004; Segal *et al.*, 2003) and Gaussian graphical modeling (Schaffer and Strimmer, 2005; Wille *et al.*, 2004; Dobra *et al.*, 2004; Li and Gui, 2006). The goal of such probabilistic modeling is to investigate the patterns of association in order to generate biological insights plausibly related to underlying biological and regulatory pathways. It is important to note that the interaction between two genes in a gene network defined by such graphical models does not necessarily imply a physical interaction, but can refer to an indirect regulation via proteins, metabolites and ncRNA that have been measured directly and therefore its interpretation depends on the model formulations (Bansal *et al.*, 2007). In this paper, we will present some details on Gaussian graphical models and methods for estimating the graphical structure in the high-dimensional settings.

It is now understood that much of a cell's activity is organized as a network of interacting modules. Such a module consists of genes co-regulated by a set of regulators to respond to different conditions (Segal *et al.*, 2003; Ernst *et al.*, 2007). It is therefore important to identify such regulatory modules. Genome-wide expression profiles provide important information on cellular states and cells' activities and therefore provide information for inferences of regulatory modules. Segal *et al.* (2003) present a probabilistic method for identifying regulatory modules from gene expression data using classification and regression tree (CART) methods. In this approach, a set of regulators including transcriptional factors and signaling proteins is first identified from literature. The model further assumes that both regulators and its targets must be regulated at the transcriptional levels, resulting in detectable changes in expression. These regulators can then be used as predictors for gene expression levels using CART. The genes that are regulated by the same set of regulators are then identified as regulatory modules. Bonneau *et al.* (2007) recently proposed a similar framework for learning parsimonious regulatory networks and called the method "Inferelator." The method first clusters genes into groups and then essentially perform linear regression with the least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996) to select the regulators that are related to the expression variations of the cluster of genes. In this paper, we will present a sparse regression mixture modeling (SRMM) approach for identifying the gene regulatory modules. Since we expect only a small set of regulators that control the expression of a module, the regression model should therefore be sparse. Our approach is based on a combination of finite mixture modeling (McLachlan and Basford, 1988) and Lasso.

Another area of intensive research in recent years has been to integrate gene expression data with motif or ChIP-chip data in order to identify transcriptional networks (Bussemaker *et al.* 2001; Keles *et al.*, 2002; Gao *et al.*, 2004, Conlon *et al.*, 2003). The foundamental idea of these approaches is based on linear regression analysis with gene expression levels as responses and motifs or ChIP-chip data as predictors. While these approaches work reasonably well in discovery of regulatory motifs in lower organisms, they often fail to identify mammalian transcriptional factor binding sites (Das *et al.*, 2006). Das *et al.* (2006) proposed to correlate the binding strength of motifs with expression levels using the multivariate adaptive smoothing splines (MARS) of Friedman (2001). Ernst *et al.* (2007) proposed an interesting approach based on the Hidden Markov model for reconstructing dynamic regulatory networks using microarray time-course gene expression data and ChIP-chip data. We will present an approach based on varying coefficient models in order to identify the transcriptional factors that are involved in a given biological process.

In this paper, we present the statistical formulations of the problems related to inference

of genetic networks based on gene expression data, inference of regulatory modules based on both gene expression data and genome annotation including information on transcriptional factors and regulators, and inference of regulatory networks based on gene expression and sequence motif or ChIP-chip data. We review some statistical methods developed for these methods and focus on the approaches that we have developed. We illustrate these methods by presenting results from analysis of several real data sets. Finally, we present a brief discussion on future work in this important area.

2 Network Inference Based on Gaussian Graphical Models

Graphical models use graphs to represent dependencies between stochastic variables (Edwards, 2000). The graphical approach yields dependence models that are easily visualized and presented. One specific graphical model is the Gaussian graphical model, which assumes that the multivariate vector follows a multivariate normal distribution with a particular structure of the inverse of the covariance matrix, often called the precision or concentration matrix. For such Gaussian graphical models, it is usually assumed that the patterns of variation in expression for a given gene will be predicted by those of a small subset of other genes. This assumption leads to sparsity (i.e., many zeros) in the precision matrix of the multivariate distribution and reduces the problem to well-known neighborhood selection or covariance selection problems (Dempster, 1970). In such a concentration graph modeling framework, the key idea is to use partial correlation as a measure of independence of any two genes, rendering it straightforward to distinguish direct from indirect interactions. This is in contrast to the covariance graphical model where marginal correlations are used. It has been demonstrated in the literature that many biochemical and genetic networks are not fully connected (Tegner *et al.*, 2003; Jeong *et al.*, 2001; Gardner *et al.*, 2003) and many genetic interaction networks contain many genes with few interactions and a few genes with many interactions. Therefore, the genetic networks are intrinsically sparse and the corresponding precision matrix should be sparse.

There are several approaches in the literature to covariance selection problems in the context of microarray data analysis. Schafer and Strimmer (2005) proposed a naive approach to estimate the precision matrix by using a boosted G-inverse, then determining which off-diagonal elements are zero by a thresholding and false discovery procedure. The drawback of this approach is that the sparsity is not accounted for when estimating the precision matrix, so the procedure is expected to perform poorly. Meinshausen and Buhlmann (2006) proposed a gene-by-gene approach by using the Lasso (Tibshirani, 1996) to find neighbors for each gene. Under a large set of assumptions they showed that the neighbors can be consistently identified when the sample size goes to infinity, which is very rare for microarray gene expression data. Dobra *et al.* (2004) proposed a Bayesian approach by converting the dependency networks into compositional networks using Cholesky decomposition. The graphs are then used to estimate the precision matrix. Since Cholesky decomposition of the precision matrix naturally imposes ordering restriction of the variables, the procedure is computationally quite intensive since it has to determine gene order in their model construction. Finally, Wille *et al.* (2004) proposed to infer Gaussian graphs based on tri-graphs by considering all partial correlations conditioning on only one other variable. Strictly speaking, the result-

ing tri-graphs are not true Gaussian concentration graphs. In the following, we provide a brief introduction ofto Gaussian graphical models and review the threshold gradient descent (TGD) approach for identifying such graphs developed in Li and Gui (2006).

2.1 Gaussian graphical models

We assume that the gene expression data observed are randomly sampled observational or experimental data from a multivariate normal probability model. Specifically, let X be a random normal p-dimensional vector and $X_1, \cdots, X_p$ denote the p elements, where p is the number of genes. Let $V = \{1, \cdots, p\}$ be the set of nodes (genes), and $X^{(k)}$ be the vector of gene expression levels for the kth sample. We assume that

$$X \sim N_p(0, \Sigma) \tag{1}$$

with positive definite variance-covariance matrix $\Sigma = \{\sigma_{ij}\}$ and precision matrix $\Omega = \Sigma^{-1} = \{\omega_{ij}\}$. This model can also be summarized as a graph model. Let $G = (V, E)$ be an undirected graph with vertex set $V = \{1, \cdots, p\}$ and edge set $E = \{e_{ij}\}$, where e_{ij}=1 or 0 according to whether vertices i and j, $1 \leq i < j \leq p$, are adjacent in G or not. The Gaussian graphical model consists of all p-variate normal distributions $N_p(0, \Sigma)$ where Σ is unknown but where the precision matrix satisfies the following linear restrictions:

$$e_{ij} = 0 \Rightarrow \omega_{ij} = 0.$$

This model is also called a covariance selection model (Dempster, 1970) or a Gaussian precision graph model.

Let $[-i]$ denote the set $\{1, 2, \cdots, i-1, i+1, \cdots, p\}$. In the Gaussian graphical model, it is well known that the partial regression coefficients of X_i on X_j in the normal linear regression $p(X_i|X_{[-i]})$ is $-\omega_{ij}/\omega_{ii}$, $j \in [-i]$, and the ijth partial correlation between the ith and the jth gene is $\rho_{ij} = -\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}$. For a given gene g, we define the neighbor of this gene as

$$ne_g = \{j : \omega_{gj} \neq 0, j \in [-g]\},$$

which contains all the genes with a non-zero partial correlation with the gene g. From the multivariate normal distribution theory, we have the following conditional independence result,

$$X_g \perp X_{G \setminus (ne_g \bigcup g)} | X_{ne_g}.$$

2.2 Threshold gradient descent regularization

We consider the estimation of the precision matrix Ω based on a sample of *i.i.d.* observations $X^{(k)} \in R^p$, $k \in N = \{1, \cdots, n\}$, where the set N can be interpreted as indexing the samples on which we observe the variables in V and $X^{(k)}$ is the kth observation. Li and Gui (2006) developed a penalized procedure for estimating Ω using the idea of threshold gradient descent (Friedman and Popescu, 2004; Gui and Li, 2005) to take into account the sparse nature of the precision matrix for genetic networks.

In order to utilize the sparse property of the precision matrix, we propose in this section to maximize the likelihood function based on model (1) subject to constraint by "sparse" precision matrix Ω. Let $\omega^d \equiv \{\omega_{11}, \cdots, \omega_{pp}\}$ denote the vector of the diagonal elements of

the matrix Ω and $\omega^o \equiv \{\omega_{ij}\}_{i \neq j}$ denote the vector of $q = p(p-1)/2$ off-diagonal elements of the Ω matrix. The likelihood function can be written as

$$w(\omega^d, \omega^o) = \frac{n}{2} \log |\Omega| - \frac{1}{2} \sum_{k=1}^{n} X^{(k)'} \Omega X^{(k)}, \tag{2}$$

where $X^{(k)}$ is the kth observation. We assume that the variables are standardized. When $p < n$, the maximum likelihood estimate (MLE) of Ω is simply the inverse of the sample covariance matrix, and when $n < p$, the MLE of Ω is not unique.

In order to account for sparsity of the precision matrix Ω, Li and Gui (2006) defined a loss function as the negative of the log likelihood function (2),

$$l(\omega^d, \omega^o) = -w(\omega^d, \omega^o).$$

Based on equation (2), the gradient of the loss function with respect to Ω is

$$\frac{\partial l}{\partial \Omega} = \frac{n}{2} \Omega^{-1} - \frac{1}{2} \sum_{k=1}^{n} X^{(k)} X^{(k)'}. \tag{3}$$

From this we can obtain the gradient of the loss function over the off-diagonal elements ω^o. Define $g(\omega^o) = (g_1(\omega^o), \cdots, g_q(\omega^o)) = -\nabla_{\omega^o} l(\omega^o, \omega^d)$ to be the negative gradient of l with respect to ω^o. To find an optimal path from all the paths from $\Omega = I$ to the MLE of Ω or to a precision matrix surface formed by $\Omega = S^-$ when $p > n$, we start from $\nu = 0$, $\omega^o = (0, \cdots, 0)$ and $\omega^d = (1, \cdots, 1)$ and update the elements ω^o by the following gradient descent step,

$$\hat{\omega^o}(\nu + \Delta\nu) = \hat{\omega^o}(\nu) + \Delta\nu h(\nu),$$

where $\hat{\omega^o}(\nu)$ is the ω^o value corresponding to current ν, $\Delta\nu > 0$ is an infinitesimal increment and $h(\nu)$ is the direction in the parameter space tangent to the path evaluated at $\hat{\omega^o}(\nu)$. This tangent vector at each step represents a descent direction. In order to direct the path towards parameter points with diverse values, following Friedman and Popescu (2004), we define $h(\nu)$ as

$$h(\nu) = \{f_j(\nu).g_j(\nu), j = 1, \cdots, q\},$$

where

$$f_j(\nu) = I[|g_j(\nu)| \geq \tau \cdot max_{1 \leq k \leq q} |g_k(\nu)|],$$

where $I[.]$ is an indicator function, and $0 \leq \tau \leq 1$ is a threshold parameter that regulates the diversity of the values of $f_j(\nu)$; larger values of τ lead to more diversity. $g(\nu)$ is the negative gradient evaluated at $\hat{\omega^o}(\nu)$ and current ω^d. Therefore, τ is the parameter that controls the degree of penalty and sparsity in the ω^o, with $\tau = 1$ giving the sparsest graphs. Instead of moving along the true gradient direction, the threshold gradient update only moves along those elements with large values of the gradient. After ω^o is updated, we update the diagonal elements of Ω, ω^d, by maximizing the log-likelihood function (2) with ω^o fixed at the current values, $\hat{\omega^o}$. This is done by using Newton-Raphson iterations.

In summary, for any threshold value $0 \leq \tau \leq 1$, the threshold gradient descent regularization algorithm for the sparse Gaussian graphical model involves the following six steps,

1. Set $\omega^o(0) = 0$, $\omega^d(0) = 1$, $\nu = 0$.
2. Calculate $g(\nu) = -\partial l/\partial \omega^o$ for the current ω^o and ω^d.
3. Calculate $f_j(\nu) = I[|g_j(\nu)| \geq \tau \cdot max_{1 \leq k \leq q}|g_k(\nu)|]$ and $h(\nu)$.
4. Update $\omega^o(\nu + \triangle\nu) = \omega^o(\nu) + \triangle\nu \cdot h(\nu)$, $\nu = \nu + \triangle\nu$.
5. Update parameters ω^d by maximizing the log-likelihood using Newton-Raphson iterations with ω^o fixed at $\omega^o(\nu + \Delta\nu)$.
6. Repeat steps 2-5.

For a given τ, it is easy to see that the likelihood function increases as the iterations increase, and different τ correspond to different paths for Ω from I to S^-. It should be emphasized that for a given τ, the threshold gradient iterations stop before it reaches S^- and the number of gradient iterations at which to stop the algorithm can be determined by cross-validation (see Section 2.3). Li and Gui (2006) particularly considered the algorithm with $\tau = 1$, which corresponds to the sparsest graph for a given TGD step, and called the proposed procedure the direct threshold gradient descent procedure. Such a procedure is expected to perform better for gene expression data since most biological or genetic networks are expected to be very sparse (Barabasi and Oltvai, 2004).

2.3 Model selection by cross-validation and bootstrap

As the iterations continue, more and more non-zero elements are selected in the precision matrix and the corresponding undirected graphs grow larger. The final model should provide the best balance between coverage (correctly identified connections/total true connections) and false-positives (incorrectly identified connections/total identified connections) (Gadner *et al.*, 2004). Li and Gui (2006) proposed to use K-fold cross-validation for choosing the number of TGD iterations, ν, where for each ν, the K-fold cross-validated log-likelihood criterion is defined as

$$CV(\nu) = \frac{1}{K}\sum_{k=1}^{K}\left(-n_k \log|\Omega_{-k}| + \sum_{i \in V_k} X^{(i)}\Omega X^{(i)}\right),$$

where n_k is the size of the kth validation set V_k and Ω_{-k} is the TGD estimate of the precision matrix based on sample $V \backslash V_k$ evaluated at $\hat{\Omega}(\nu)$. Alternatively, we can use the BIC criteria for selecting ν, where the degrees of freedom can be defined as the number of nonzero entries of the off-diagonal elements of the precision matrix. This is similar in spirit to Lasso in linear regression where the degrees of freedom is defined as the number of nonzero coefficients (Zou *et al.*, 2007).

Since the number of the off-diagonal elements in the precision matrix is often quite large compared to the sample size, there is often considerable uncertainty in the edges chosen. As a final step in the procedure, we propose to use the bootstrap method to determine the statistical accuracy and the importance of each of the edges identified by the TGD procedure. In bootstrapping, B bootstrap data sets, $X^{*1}, \cdots, X^{*B}$, are sampled with replacement from the original data set such that each bootstrap sample contains n observations. We then

apply the TGD procedure to each bootstrap data set and examine which edges are in the final models. One can then choose only the edges with high probability of being non-zero in the precision matrix over the bootstrap samples.

2.4 Simulation results and application to real data set

Li and Gui (2006) performed simulation studies to evaluate the proposed threshold gradient descent procedure and applied this to analysis of isoprenoid metabolic pathways. Results indicate that by accounting for sparsity in estimating the precision matrix, one can obtain a better estimate of the precision matrix, and the TDG procedure can effectively identify the linked edges in the Gaussian graphical models.

Li and Gui (2006) applied the TGD procedure to analysis of the Arabidopsis thaliana isoprenoid pathway. The isoprenoid biosynthetic pathway provides intermediates of many natural products including steroles, chlorophylls, carotenoids, plastoquinone and abscisic acid, etc. It is now known that plants contain two pathways for the synthesis of the structural precursors of isoprenoids: the mevalonate (MVA) pathway, located in the cytosol/ER, and the recently discovered methylerythritol 4-phosphate (MEP) pathway, located in the plastids. The pathway in plastids, which is mevalonate-independent, occurs and is responsible for the subsequent biosynthesis of plastidiar terpenoids such as carotenoids and the side chains of chlorophyll and plastoquinone (Wille *et al.*, 2004). It is therefore important to understand the organization and regulation of this complex metabolic pathway, with the long-term goal of using the generated knowledge to undertake metabolic engineering strategies oriented to increase the production of isoprenoids with pharmaceutical and food applications, and also to the design and development of new antibiotics.

In order to better understand the pathway and gain insights into the crosslink between the two pathways at the transcriptional level, Wille *et al.* (2004) reported a data set including the gene expression patterns monitored under various experimental conditions using 118 GeneChip microarrays. For the construction of the genetic network, they focused on 40 genes, 16 of which were assigned to the cytosolic MVA pathway, 19 to the plastidal MEP pathway and five genes encoding proteins located in the mitochondria. See the solid lines of Figure 1 for the MVA and the MEP pathways and the genes involved.

In order to demonstrate whether the proposed TGD method can identify the known isoprenoid pathways of these 40 genes based on the 118 gene expression measurements, Li and Gui (2006) first estimated the precision matrix by the threshold gradient methods. Using 10-fold cross-validation, the TGD procedure resulted in 20 non-zero off-diagonal elements. We next used a bootstrap with the TGD procedure to estimate the confidence of the edges. With bootstrap probability of 0.50 or higher, we identified 19 pairs of genes that are connected with high confidence, of which 12 pairs have a bootstrap probability of 0.80 or higher. These 19 pairs are plotted on the true network in Figure 1. We find a module with strongly interconnected genes in each of the two pathways. For the MEP pathway, DXPS2, DXR, MCT, CMK and MECPS are connected as the true pathway. Similarly, the genes in the MVA pathways, AACT2, HMGR2, MK, MPDC1, FPPS1 and FPP2 are closely connected. In addition, there are also several genes in the MEP pathway that are linked to proteins in the mitochondria.

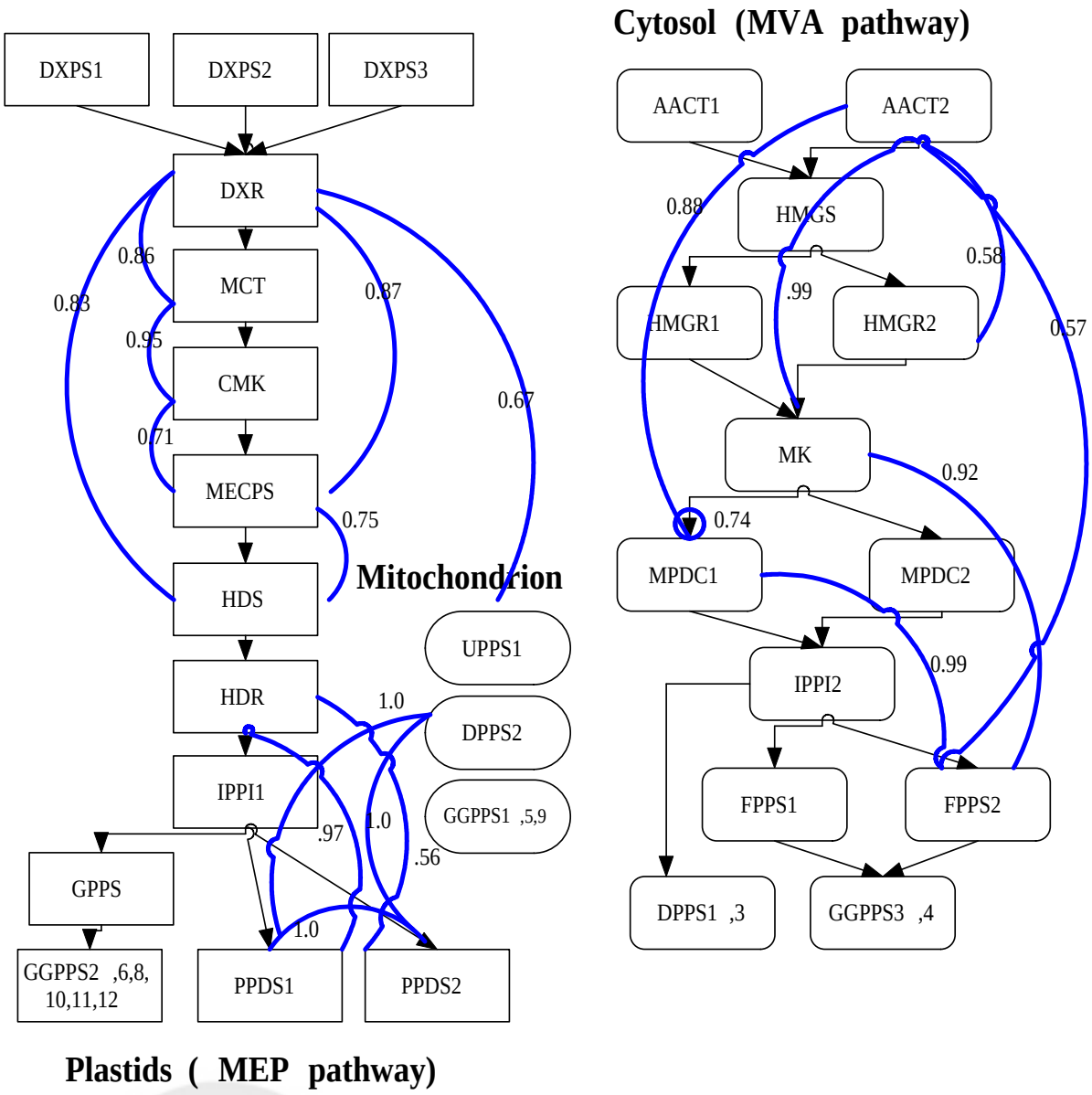

Figure 1: *Pathways identified by the TGD method for the 40 genes in the isoprenoid pathways, where the solid arrows are the true pathways, curved undirected lines are the estimated edges with bootstrap probability of greater than 0.5 for the TGD method. For this plot, the left panel represents a subgraph of the gene module in the MEP pathway and the right panel represents a subgraph of the gene module in the MVA pathway. The numbers on the estimated edges are the bootstrap probabilities.*

3 Methods for Identifying Regulatory Modules

Since the networks inferred from gene expression data alone do not imply any physical interactions among genes, it is important to incorporate other data sources to infer regulatory networks. In addition, influence interactions may include physical interactions if the two interacting partners are a transcriptional factor and its target, suggesting that incorporating the transcriptional factors information can help to identify the relevant regulatory networks and modules. Segal *et al.* (2003) proposed to integrate gene expression data and knowledge of transcriptional factors in order to identify transcriptional modules using regression trees in a mixture modeling framework. Lee *et al.* (2007) further developed this idea for integrating microarray gene expression data, SNPs data and regulatory information in order to identify the genetic variations that affect the regulatory modules. In this section, we reformulate the problem into sparse regression mixture modeling (SRMM) and propose a Lasso-EM algorithm for identifying such regulatory modules.

3.1 The SRMM for identifying transcriptional modules

Consider the microarray gene expression data setup of G genes over C experimental conditions. These G genes also include C transcriptional factors or signaling proteins. Let Y_{gc} be the log-expression level of the gth gene at the experimental condition c, for $g = 1, \cdots, G$ and $c = 1, \cdots, C$ and $Y = \{Y_{gc}, g = 1, \cdots, G, c = 1, \cdots, C\}$. Let X_{rc} be log-expression level of the rth regulator at the condition c, for $r = 1, \cdots, R$ and $c = 1, \cdots, C$, $X_c = \{X_{1c}, \cdots, X_{Rc}\}$ be the expression level of the R regulators at the cth experiment condition, and $X = \{X_{rc}, r = 1 \cdots, R, c = 1, \cdots, C\}$.

Assuming that there are K regulatory modules, let M_g be the module membership for the gene g. We assume the following model for the observed expression data for genes in the kth regulatory module,

$$\begin{aligned} Y_{gc}|\{M_g = k\} &= \sum_{r=1}^{R} X_{rc}\beta_{kr} + \epsilon_{gc}, \\ M_g &\sim Multinomial(\pi), \\ \text{with } \pi &= (\pi_1, \cdots, \pi_K)^T, \pi_k \leq 0, \text{and } \sum_{k=1}^{K} = 1, \end{aligned} \tag{4}$$

where $\beta_k = (\beta_{k1}, \cdots, \beta_{kR})^T$ is the vector of module-specific parameters, ϵ_{gc} is the error term, which is assumed to follow a $N(0, \sigma_k^2)$, and π_i is the prior probability that a gene belong to the ith module. Since the expression level of a given gene is often regulated by a small set of regulators, for a given k, we should expect that many of the elements in vector β_k should be zero. We call model (4) the SRMM model. In this mixture model formulation, the unknown parameters include K, β_k and σ_k^2 for $k = 1, \cdots, K$ and π. Note that the SRMM model (4) can be easily extended to include interaction terms between the regulators,

$$Y_{gc}|\{M_g = k\} = \sum_{r=1}^{R} X_{rc}\beta_{kr} + \sum_{r,r'} X_{rc}X_{r'c}\beta_{krr'} + \epsilon_{gc},$$

where $\beta_{krr'}$ is used to model the interaction effect between the regulators r and r' on gene expression levels for genes in the kth module.

3.2 An EM algorithm based on Lasso

In order to deal with the problem of a large R and to account for the sparse nature of the parameters β_k in model (4), we propose to develop the following EM algorithm based on the Lasso for estimating the model parameters. Let $M = \{M_{g1}, \cdots, M_{gK}, g = 1, \cdots, G\}$ be the matrix of module-membership indicators for the G genes, where M_{gk} is 1 if the gth gene belongs to the kth module. The complete data log-likelihood can be written as

$$l(\beta_k, \sigma_k^2 \pi; Y, M|X) \;=\; l_1(M;\pi) + l_2(Y|M, X; \beta_k, \sigma_k^2),$$

where

$$\begin{aligned} l_1(M;\pi) &= \sum_{g=1}^{G}\sum_{k=1}^{K} M_{gk}\log(\pi_k), \\ l_2(Y|M, X; \beta_k, \sigma_k^2) &= -\frac{1}{2}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{c=1}^{C} M_{gk}\left\{\log(\sigma_k^2) + \frac{(Y_{gc} - X_c^T\beta_k)^2}{\sigma_k^2}\right\}. \end{aligned} \tag{5}$$

It is easy to show that in the $(t+1)$th E-step, we have

$$\hat{M}_{gk} = E(M_{gk} = 1|X, Y) = \frac{\pi_k^{(t)} Pr(Y_g|M_{gk} = 1; \beta_k^{(t)}, \sigma_k^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} Pr(Y_g|M_{gk} = 1; \beta_k^{(t)}, \sigma_k^{(t)})},$$

where $Pr(Y_g|M_{gk} = 1; \beta_k^{(t)}, \sigma_k^{(t)})$ is the normal density for Y_g if the gth gene belongs to the kth module. At the $(t+1)$th M-step, it is easy to check that the EM equation for updating the prior probability is given by

$$\hat{\pi}_k^{(t+1)} = \frac{\sum_{g=1}^{G} \hat{M}_{gk}}{G}.$$

However, the M-step for updating the parameter β_k needs to account for the sparsity of this parameter. From the expression of l_2 in equation (5), we need to find the β_k that minimizes the following quantity,

$$\sum_{g=1}^{G} \hat{M}_{gk} \sum_{c=1}^{C} (Y_{gc} - X_c^T\beta_k)^2 = \sum_{g=1}^{G}\sum_{c=1}^{C} (\sqrt{\hat{M}_{gk}} Y_{gc} - \sqrt{\hat{M}_{gk}} X_c^T\beta_k)^2,$$

subject to sparsity constraint of β_k,

$$|\beta_k|_1 = \sum_{r=1}^{R} |\beta_{kr}| < s,$$

where s is a tuning parameter, assumed to be the same for all $k = 1, \cdots, K$. This is equivalent to performing linear regression with $\sqrt{\hat{M}_{gk}} Y_{gi}$ as responses and $\sqrt{\hat{M}_{gk}} X_{gi}^T$ as regressors. We can then use Lasso to update β_k. However, since $G \times C$ is often very large, such an implementation can be very time-consuming. Alternatively, we can use a sparse version of the EM algorithm or the Hard-EM algorithm to update β_k. Specifically, we first

classify the gene g into the k_gth module, where $k_g = \text{argmax}_k \hat{M}_{gk}$, for $g = 1, \cdots, G$. We then estimate β_k using Lasso based on the data of the genes in the current kth module under the constraint that $|\beta_k|_1 < s$. This can be efficiently implemented using the R Lars function (Efron *et al.*, 2004).

Finally, after obtaining the update of β_k, we can update the error variance by

$$\hat{\sigma}_k^{2(t+1)} = \frac{\{\sum_{g=1}^{G} \hat{M}_{gk}(Y_g - X^T\beta_k^{(t+1)})^T(Y_g - X^T\beta_k^{(t+1)})\}}{\sum_{g=1}^{G} C\hat{M}_{gk}}.$$

We call this EM algorithm using Lasso the Lasso-EM algorithm. After the convergence of the Lasso-EM algorithm, we can partition the genes into K regulatory modules. Specifically, we partition gene g into module k if $k = argmax_{k'}\{Pr(M_g = k')\}$. In addition, we can obtain the regulation program of a module specifying the set of regulatory genes in the module that controls the module and the mRNA expression profile of the genes in the module as a function of the expression of the module's regulators. Specifically, for the kth module, the regulation program includes the regulator set $\{r : \hat{\beta}_{kr} \neq 0\}$ and the sign and the magnitude of $\hat{\beta}_{kr}$ determine how the regulation program controls the expression of the module. Note that our method allows some regulators to participate in the regulation programs of multiple modules and also allows a group of genes that is not regulated by any of the regulators.

3.3 Selection of the number of modules K and the tuning parameter s

In order to implement the proposed Lasso-EM algorithm, for each given number of the modules K, we need to determine the tuning parameter $s(K)$. We also need to determine the number of the regulatory modules K. As commonly used for mixture models, for a given number of regulatory modules k, we can choose the tuning parameter s by maximizing the BIC score, which is defined as

$$BIC(s(k)) = \text{l(s(k))} - p(s(k)) \log(G \times C)$$

for the model with k cluster and tuning parameter s, where $p(s(k)) = \sum_{k=1}^{k}\sum_{r=1}^{R} I(\beta_{kr} \neq 0)$ is the total number of non-zero parameters in the model (Zou *et al.*, 2007) and $G \times C$ is the number of observations, and the log-likelihood for the model with k clusters and tuning parameter s can be written as

$$l(s(k)) = \log \sum_{g=1}^{G}\sum_{k=1}^{K} \pi_k (2\pi\sigma_k^2)^{-C/2} \exp\left\{-\frac{\sum_{c=1}^{C}(Y_{gc} - X_c^T\beta_k)^2}{\sigma_k^2}\right\}.$$

We then choose K as $K = \text{argmax}_k\ l(s(k))$.

3.4 Application to yeast stress data set

To demonstrate the proposed Lasso-EM algorithm for identifying the transcriptional modules, we applied the proposed method to analysis of the yeast data set reported in Segal *et al.* (2003), consisting of 2,355 genes, 466 candidate regulators (transcriptional factors and signaling proteins) and 173 arrays of the yeast stress data, measuring gene expression responses to

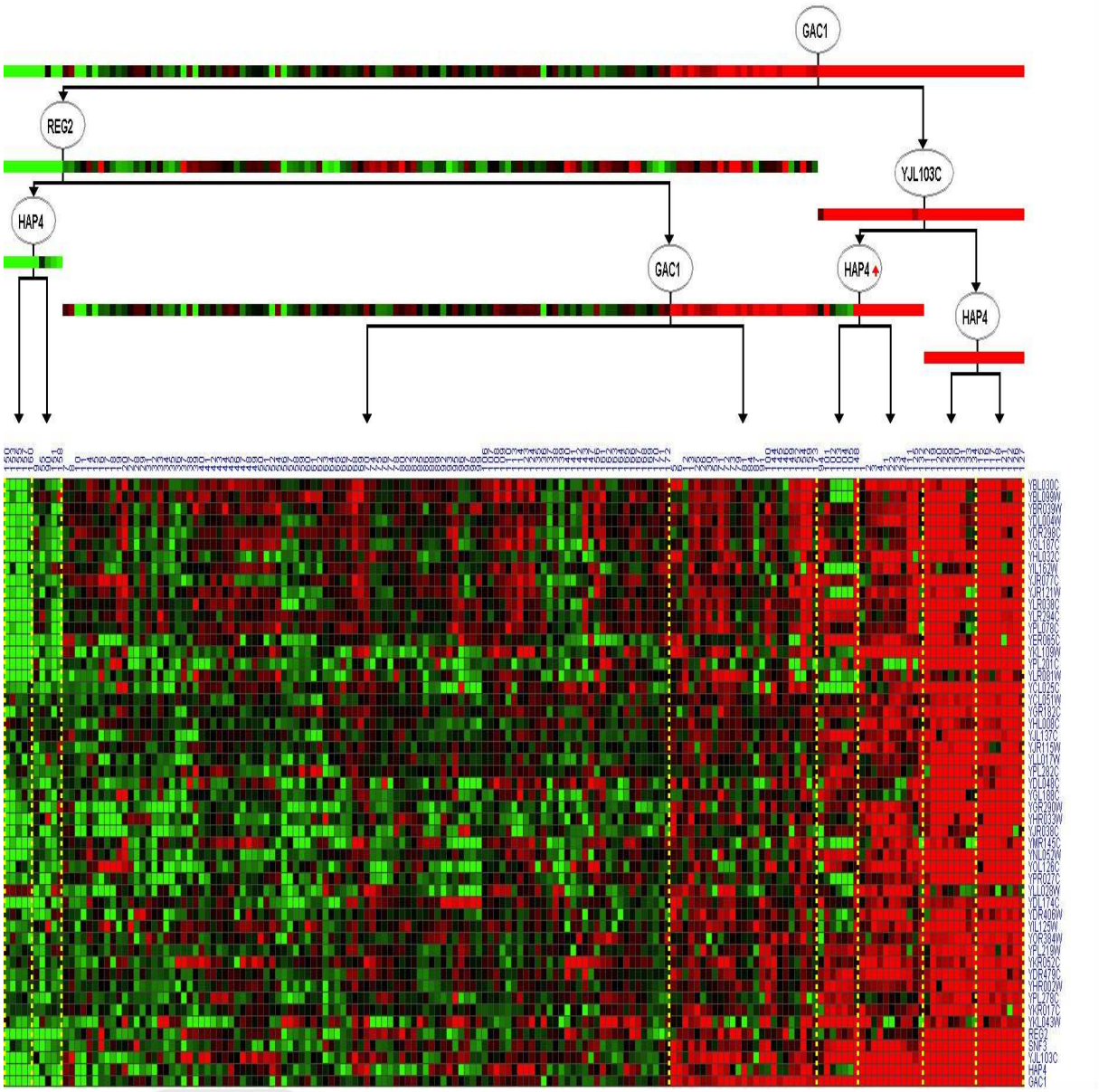

Figure 2: *The respiration and carbon module identified from analysis of the yeast stress responses microarray gene expression data set using the proposed Lasso-EM algorithm. The plot was generated using the plotting tool provided in Segal et al. (2003). The heatmap is for gene expression, where genes are plotted as rows and arrays are columns. Arrays are arranged according to the regulation tree, where each node represents a regulator and the expression of the regulators is shown below their respective node.*

various stress conditions. This set of 2,355 genes has a significant change in gene expression at the measured stress conditions, excluding members of the generic environmental stress response cluster.

We started the Lasso-EM algorithm based on results from the hierarchical clustering. Allowing for at most 5 regulators for each module, the Lasso-EM algorithm identifie35 d regulatory modules, each containing several significant GO processes, indicating that the same set of regulators may regulate different biological processes. As an example, Figure 2 shows a heatmap for the Respiration/Carbohydrate metabolism module, including four regulators, Hap4, GAC1, Reg2 and SNF.3 (YJL103C), and their effects on gene expression levels for genes in this module are given by the following linear model,

$$y = 0.12Hap4 + 0.015GAC1 + 0.047Reg2 + 0.06SNF3.$$

This model specified the Hap4 transcriptional factor as an important activating regulator, consistent with the known role of Hap4 in activation of respiration (Segal *et al.*, 2003). This model also suggests that the protein phosphatase type 1 regulatory subunit Gac1 and transcription factor Reg2 may also regulate the expression levels of this module. The results largely agree with the similar module identified by Segal *et al.* (2003).

4 Inference of Transcriptional Networks

Since many essential biological systems or processes are dynamic systems, it is important to study the gene expression patterns over time in a genomic scale in order to capture the dynamic behavior of gene expression. Research in analysis of such microarray time-course (MTC) gene expression data has focused on two areas: clustering of MTC expression data (Luan and Li, 2003; Ma *et al.*, 2006) and identifying genes that are temporally differentially expressed (Hong and Li, 2006; Yuan and Kendziorski, 2006; Tai and Speed, 2006; Storey *et al.*, 2005). While both problems are important and biologically relevant, they provide little information about our understanding of gene regulations. We present in this section the methods based on the "physical interaction" approach that aim to identify interactions among transcriptional factors and their target genes through sequence motifs and ChIP-chip data found in promotor sequences. In particular, Wang *et al.* (2007) considered integrating microarray time-course gene expression data and motif or ChIP-chip data in order to identify the transcriptional factors that are involved in gene expression variations during a given biological process.

4.1 Functional response model with time-varying coefficients for MTC gene expression data

We consider a microarray time-course gene expression experiment. Let $Y_i(t)$ be the expression level of the ith gene at time t, for $i = 1, \cdots, n$. We assume the following regression model with functional response,

$$Y_i(t) = \mu(t) + \sum_{k=1}^{K} \beta_k(t) X_{ik} + \epsilon_i(t), \tag{6}$$

where $\mu(t)$ is the overall mean effect, $\beta_k(t)$ is the regulation effect associated with the kth transcriptional factor, X_{ik} is the matching score or the binding probability of the kth transcriptional factor on the promoter region of the ith gene, and $\epsilon_i(t)$ is a realization of a zero-mean stochastic process. Several different ways and data sources can be used to derive the matching score X_{ik}. One approach is to derive the score using the position-specific weight matrix (PSWM). Specifically, for each candidate TF k, let P_k be the positive specific weight matrix of length L, b with element $P_{kj}(b)$ being the probability of observing the base b at position j. Then each L-mer l in the promoter sequence of the ith gene is assigned a score S_{ikl} as:

$$S_{ikl} = \sum_{j=1}^{L} \log \frac{P_{kj}(b_{ilj})}{B(b_{ilj})},$$

where b_{ilj} is the nucleotide at position j on the lth sequence for gene i, and $B(b)$ is the probability of observing b in the background sequence. This score always assumes a value between 0 and 1. We then define $X_{ik} = max_l S_{ikl}$, which is the maximum of the matching scores over all the L-mer in the promoter region of the ith gene. The maximum scores can then be converted into the binding probabilities using the method described in Chen *et al.* (2007). Alternatively, we can define the binding probability based on the chromatin immunoprecipitation (ChIP-chip) data (Wang *et al.*, 2007; Chen *et al.*, 2007).

4.2 Estimation using B-splines

We consider estimation of the nonparametric function in Model (6) using the smoothing spline method by approximating $\beta_k(t)$ by using the natural cubic B-spline basis,

$$\beta_k(t) = \sum_{l=1}^{L+4} \beta_{kl} B_l(t) \tag{7}$$

where $B_l(t)$ is the natural cubic B-spline basis function, for $l = 1, \cdots, L+4$, where L is the number of interior knots. Replacing $\beta_k(t)$ by its B-spline approximation in equation (7), Model (6) can be approximated as

$$Y_i(t) = \mu(t) + \sum_{k=1}^{K} \left\{ \sum_{l=1}^{L+4} \beta_{kl} [B_l(t) X_{ik}] \right\} + \epsilon_i(t), \tag{8}$$

where we have K groups of parameters, with $\beta_k^* = \{\beta_{k1}, \cdots, \beta_{kL+4}\}$ being the parameters associated with the group k, and we want to select the groups with non-zero coefficients. This is the grouped variable selection problem considered in Yuan and Lin (2006).

4.3 A group SCAD penalization procedure

Wang *et al.* (2007) proposed a general group SCAD (gSCAD) procedure for selecting the groups of variables in a linear regression setting. Selecting important variables in Model (6) corresponds to the selection of groups of basis functions in Model (8). Yuan and Lin (2006) proposed several procedures for such group variable selection, including group LARS and group Lasso. Instead of using the L_1 penalty for group selection as in Yuan and Lin (2006),

we propose to use the SCAD penalty of Fan and Li (2001). Specifically, to select non-zero $\beta_k(t)$, we can minimize the following penalized loss function

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n}\sum_{j=1}^{T}[y_{ij} - \mu(t_j) - \sum_{k=1}^{K}\sum_{l=1}^{L+4}\beta_{kl}B_l(t_j)X_{ik}]^2 \\ &\quad + nT\sum_{k=1}^{K} p_\lambda(||\beta_k^*||_2), \end{aligned} \tag{9}$$

where y_{ij} is the observed gene expression level for gene i at time t_j, $p_\lambda(.)$ is the SCAD penalty with λ as a tuning parameter, which is defined as

$$p_\lambda(|w|) = \begin{cases} \lambda|w| & \text{if } |w| \le \lambda, , \\ -\frac{(|w|^2 - 2a\lambda|w| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |w| < a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |w| > a\lambda \end{cases} \tag{10}$$

and $||\beta_k^*||_2 = \sqrt{\sum_{l=1}^{L+4}\beta_{kl}^2}$. The penalty function (10) is a quadratic spline function with two knots at λ and $a\lambda$, where a is another tuning parameter. Fan and Li (2001) showed that the Bayes risks are not sensitive to the choice of a and suggested using $a = 3.7$.

4.4 Numerical algorithm, properties and application

Wang *et al.* (2007) proposed a local quadratic approximation algorithm, similar to that in Fan and Li (2001), to perform the optimization problem of equation (9) and developed a GCV procedure for selecting the tuning parameter λ in the SCAD penalty function (10). They also established the oracle property of the gSCAD estimates. Wang *et al.* (2007) also performed simulation studies to evaluate the gSCAD procedure. Simulation results indicate that for a similar false positive rate, the gSCAD procedure is more sensitive in identifying the relevant transcriptional factors than simple linear regression. In addition, the estimates of the transcriptional effects are less variable than those obtained from simple linear regression analysis.

We applied the proposed methods to the analysis of cell cycle MTC data corrected by Spellman *et al.* (1998). The cell cycle is one of life's most important processes, and the identification of cell cycle regulated genes has greatly facilitated the understanding of this important process. Spellman *et al.* (1998) monitored genome-wide mRNA levels for 6178 yeast ORFs simultaneously using several different methods of synchronization including an α-factor-mediated G_1 arrest, which covers approximately two cell-cycle periods with measurements at 7-min intervals for 119 mins with a total of 18 time points (http://genome-www.stanford.edu/cellcycle/data/rawdata/). Using data based on different synchronization experiments, Spellman *et al.* (1998) identified a total of about 800 cell cycle regulated genes, some showing periodic expression patterns only in a specific experiment. Using a model-based approach, Luan and Li (2003) identified 297 cell-cycle regulated genes based on the α-factor synchronization experiments. We applied the mixture model approach described in previous section using the ChIP data of Lee *et al.* (2002) to derive the binding probabilities X_{ik} for these 297 cell cycle-regulated genes for a total of 96 transcriptional factors with at least one nonzero binding probability in the 297 genes.

We applied the gSCAD procedure with $L = 2$ and an additional L_2 penalty in order to identify the TFs that affect the expression changes over time for these 297 cell cycle regulated genes in the α-factor synchronization experiment. The gSCAD procedure identified a total of 71 TFs that are related to yeast cell cycle processes, including 19 of the 21 known and experimentally verified cell cycle-related TFs. The estimated transcriptional effects of these 21 TFs are shown in Figure 3, except for the two TFs (CBF1 and GCN4) that were not selected by the gSCAD procedure and the TF LEU3, the other 18 TFs all showed time-dependent effects of these TFs on gene expression levels. In addition, the effects followed similar trends between the two cell cycle periods. It was not clear why CBF1 and GCN4 were not selected by the gSCAD. The minimum p-values over 18 times points from simple linear regressions are 0.06 and 0.14, respectively, also indicating that CBF1 and GCN4 were not related to expression variation over time. Overall, the model can explain 43% of the total variations of the gene expression levels.

To assess false identifications of the TFs that are related to a dynamic biological procedure, we randomly permuted the gene expression values across genes and time points and applied the gSCAD procedure again to the permuted data sets. We repeated this procedure 50 times. Among the 50 runs, 5 runs selected 4 TFs, 1 run selected 3 TFs, 16 runs selected 2 TFs and the rest of the 28 runs did not select any of the TFs, indicating that our procedure indeed selects the relevant TFs with few false positives.

5 Discussion, Conclusions and Future Research

In this paper, we have reviewed several important problems and statistical methods related to analysis of genetic networks and regulatory modules based on integrating microarray gene expression data and ChIP-chip data, including the problem of constructing genetic networks based on microarray gene expression data and Gaussian graphical models; the problem of identifying regulatory modules based on gene expression data and a pre-defined set of potential regulators, and the problem of identifying regulatory networks based on both microarray gene expression data and motif and ChIP-chip data. Our review mainly emphasizes the statistical formulation of the problems and our solutions to these problems. Applications to real data sets are also briefly discussed and presented. It should be emphasized that these algorithms only generate what can be loosely referred to as a "first approximation" to a gene regulatory network. The results of this method should not be interpreted as the definitive regulatory network but rather as a network that suggests (possibly indirect) regulatory interactions (Thorsson *et al.*, 2005; Bonneau *et al.*, 2007). It should also be noted that this paper emphasizes the use of gene expression data for inferences of genetic networks and regulatory modules; however, accurate protein-level measurements of TFs will invariably have a more direct influence on the mRNA levels of the genes they regulate.

The paper only covers three problems in inference of genetic networks and regulatory modules. As more and more large-scale genomic and epigenomic data are being generated, novel statistical and computational methods are required for many other problems related to analysis of microarray gene expression data. We present two such areas that need further methodological development.

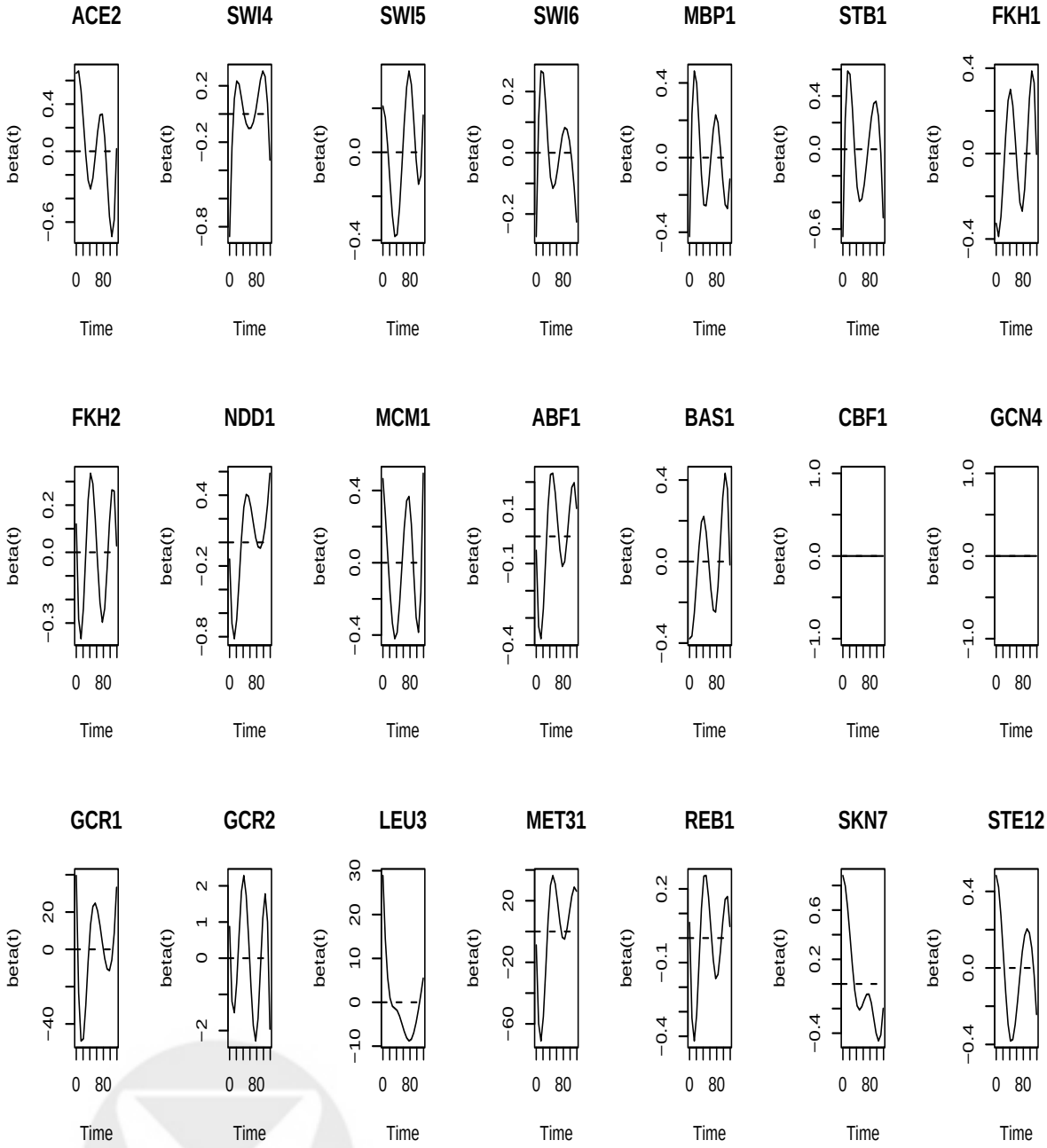

Figure 3: *Estimated time-dependent transcriptional effects for 21 known yeast transcriptional factors related to the cell cycle process using gSCAD. Note that CBF1 and GCN4 were not selected by gSCAD.*

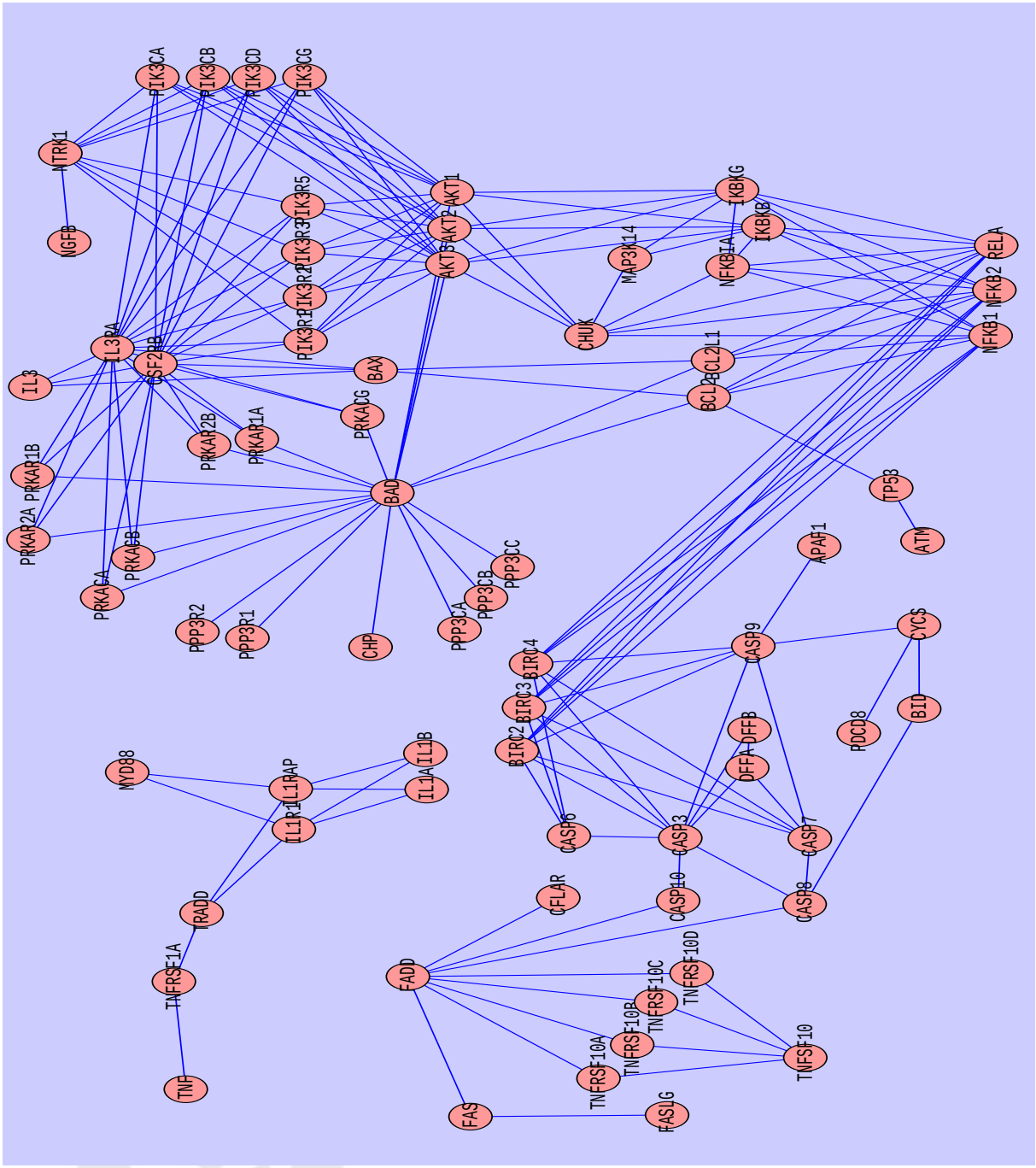

Figure 4: *KEGG Apoptosis regulatory pathway: nodes are genes and edges represent the regulatory relationship between genes. Only genes with the corresponding probe-pairs found on the Affymetrix U133A are plotted.*

5.1 Incorporating network information into analysis of microarray gene expression data

While the focus of this paper is on inference of genetic pathways and networks, an equally important problem is to incorporate the pathway information derived from data or from pathway databases into analysis of clinical phenotype data in order to identify the pathways and networks that are related to various clinical phenotypes. There is a great need for methods that can link numerical microarray gene expression data measured on the networks to the phenotypes in order to obtain biologically interpretable results. As an example, Figure 4 shows the regulatory Apoptosis KEGG pathway (Kanehisa and Goto, 2002), providing information on the regulatory relationship for genes on this pathway. We are now able to measure the gene expression levels for genes on this pathway for a sample of patients with different clinical phenotypes. We should expect certain dependency of differential expression states for genes that are neighbors on this pathway. Incorporating such prior local dependency of gene expressions into analysis of phenotype data can potentially gain power in identifying the relevant genes. Some preliminary work from our group indicates that such network-based analysis of gene expression data can greatly increase the sensitivity of identifying the relevant pathways or subnetworks (Wei and Li, 2007ab). For example, Wei and Li (2007b) developed a Markov random field (MRF) approach for identifying differentially expressed genes between two different experimental conditions, where a discrete MRF is used to model the local dependency of the differential expression states for genes on the network. They demonstrated by both simulations and analysis of several breast cancer gene expression studies that such an MRF-based methods can identify more biologically interpretable genes and sub-networks.

Novel methods are also needed to formally incorporate network information into regression models. One solution to this problem is to perform a network-imposed smoothness penalty in order to obtain locally smoothed estimates of the regression parameters. Assume that Y follows a distribution in an exponential family with mean $\mu = E(Y)$ and variance $V = Var(Y)$. The generalized linear model (GLM) (McCullagh and Nelder, 1989) models the random component of Y through a link function g:

$$g(\mu) = \sum_{k=1}^{p} X_k \beta_k, \tag{11}$$

where X_k is the gene expression measurement of the kth genes and β_k is the regression coefficient corresponding to the kth gene, for a total of p genes on the genetic network $N = (G, E)$ with gene set G and edge set E. We further denote $\beta = \{\beta_1, \cdots, \beta_p\}$ as the vector of regression parameters of the model. Suppose that we have n *i.i.d.* observations $(x_i, y_i), i = 1, \cdots, n$ of a gene expression vector x_i and a response variable y_i. We can then define an estimate of β by minimizing the following regularized loss function,

$$\hat{\beta} = argmin \left\{ l(\beta) + \lambda_1 \sum_{k=1}^{p} |\beta_k| + \lambda_2 \beta' L \beta \right\}, \tag{12}$$

where $l(\beta)$ is a loss function (e.g., negative of the log-likelihood function corresponding to the GLM (11)), L is the Laplacian matrix as defined in Chung (1997) and λ_1 and λ_2 are two

tuning parameters. It is easy to verify that

$$\beta' L\beta = \sum_{i\sim j\in E} (\beta_i/\sqrt{d_i} - \beta_j/\sqrt{d_j})^2, \tag{13}$$

where $i \sim j$ denotes that ith gene and jth gene are neighbors on the network N and d_i and d_j are the respective degrees. The scaling of the β parameters by their degrees is used to reflect the fact that the genes with more neighbors (e.g., the "hub" genes) tend to play a more important biological role and therefore should correspond to larger coefficients. In this regularized formulation (12), the first L_1 penalty leads to sparse solution and the second penalty $\beta' L\beta$ leads to a smoothness solution with respect to the network structure, i.e., it enforces that the degree-scaled β estimates are similar for genes that are neighbors on the network. In order to account for both activation and inhibition effects, we can modify the smoothness penalty (13) as

$$\lambda_2 \left\{ \sum_{i\sim j\in E^+} (\beta_i/\sqrt{d_i} - \beta_j/\sqrt{d_j})^2 + \sum_{i\sim j\in E^-} (\beta_i/\sqrt{d_i} + \beta_j/\sqrt{d_j})^2 \right\},$$

where E^+ is the set of transcriptional activation edges ($\longrightarrow$) and E^- is the set of transcriptional inhibition edges ($\dashv$). This modification is used to reflect the fact that β_i and β_j are expected to have the same sign if $(i \sim j) \in E^+$ and different signs if $(i \sim j) \in E^-$. Similar to the Elastic-Net penalty of Zou and Hastie (2005), the optimization of equation (12) can be efficiently solved by using the R Lars algorithm (Efron *et al.*, 2004) and the tuning parameters λ_1 and λ_2 can be chosen using cross-validation.

5.2 Development of statistical and computational methods for integrating gene expression data and epigenomic data

Another important area for future research is to develop rigorous statistical and computational methods for integrating microarray gene expression data with other types genomic data for an even more detailed understanding of genetic networks, especially genetic regulatory networks. We have reviewed two such approaches: one uses the transcriptional factor annotation information in inference of regulatory modules, the other uses the sequence motif and ChIP-chip data on inference of regulatory models. However, other genome-wide data can be very useful. For example, we know that sequence polymorphisms affect gene expression by perturbing the complex networks of regulatory interactions. It is therefore important to simultaneously consider both single nucleotide data and the gene expression data in order to obtain both *cis*− and *trans*− effects on gene expression (Brem *et al.*, 2005; Schadt *et al.*, 2005; Morley *et al.*, 2004). Standard methods attempt to associate each gene expression phenotype with genetic polymorphisms. Lee *et al.* recently developed an interesting method to understand the mechanisms by which genetic changes perturb gene regulation by combining SNP, gene expression and transcriptional factors information.

Statistical and computational methodologies for genomic data analysis and integration are also needed for analysis of epigenomics data, with the aim to understand systems-level gene regulatory mechanisms. A multi-cellular organism contains only one genome, but different cell types contain different epigenomic patterns, including chromatin structure (Steifeld

et al., 2007), histone modification (Yuan *et al.*, 2005), nucleosome positions (Segal *et al.*, 2006) and DNA methylations (Eckhardt *et al.*, 2006). These epigenomic markers are important for regulating protein-DNA binding activities and gene transcription. As more and more epigenomic data become available (Yuan *et al.*, 2005; Eckhardt *et al.*, 2006; Heintzman *et al.*, 2007), it is important to develop novel statistical methods for analyzing such data together with gene expression data in order to estimate the direct regulatory effects of epigenomic factors.

5.3 Final Remarks

Elucidating genetic pathways and networks is one of the most important problems in modern biology research. Microarray gene expression data together with other high-throughput genomic, proteomic and epigenomic data provide the opportunity to derive such networks and apply network knowledge to study other important biological systems such as disease initiation and progression. However, statistical and computational methods for analyzing these data are becoming even more important. We conclude this paper by quoting Dr. Collins's *Nature* article, "*computational methods have become intrinsic to modern biological research, and their importance can only increase as large-scale methods for data generation become more prominent, as the amount and complexity of the data increases and as the questions being addressed become more sophisticated* " (Collins *et al.*, 2003).

Acknowledgments

This research is supported by NIH grant ES009911 and a grant from the Pennsylvania Health Department. I thank my students Zhandong Liu and Caiyan Li for their work on SRMM; Zhi Wei for his work on the local MRF model; my postdoctoral fellows Dr. Jiang Gui for his work on the TGD procedure and Dr. Lifeng Wang for his work on gSCAD; and Edmund Weisberg, MS for editorial help. I also thank Professor Dehmer and Dr. Emmert-Streib for inviting me to contribute this paper to the book.

References

Bansal M, Belcastro V, Ambesi-Impiombato A and di Bernardo D (2007): How to infer gene networks from expression profiles. *Molecular Systems Biology*, 3: 78.

Barabasi AL and Oltvai ZN (2004): Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5: 101-113.

Bonneau R, Reiss D, Shannon P, Facciotti M, Hood L, Baliga N, Thorsson V (2006): The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biology*, 7: R36

Brem RB, Storey JD, Whittle J and Kruglyak L (2005): Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*, 436: 701703.

Bussemaker HJ, Li H and Siggia ED (2001): Regulatory element detection using correlation with expression. *Nature Genetics*, 27: 167-171.

Chen G, Jensen S, and Stockert C (2007): Clustering of genes into regulons using integrated moeling(cogrim). *Genome Biology*, 8, 1, R4.

Chung F (1997): *Spectral Graph Theory*, Vol 92 of CBMS Reginal Conferences Series. American Mathematical Society, Providence.

Collins FS, Green ED, Guttmacher AE and Guyer MS (2003): A vision for the future of genomics research. *Nature*, 422: 835 - 847.

Conlon EM, Liu XS, Lieb JD and Liu JS (2003): Integrating regulatory motif discovery and genome-wide expression analysis *Proceedings of National Academy of Sciences*, 100: 3339-3344;

Das D, Nahle Z and Zhang MQ (2006): Adaptively inferring human transcriptional subnetworks. *Molecular Systems Biology*, msb410067-E1.

Demspter AP (1972): Covariance selection. *Biometrics*, 28: 157-175.

Dobra A, Jones B, Hans C, Nevis J and West M (2004): Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90: 196-212.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S (2006): DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, 38(12): 1378-1385.

Edwards D (2000): *Introduction to Graphical Modelling*. 2nd edition. New York: Springer Verlag.

Efron B, Hastie T, Johnstone I and Tibshirani R (2004): Least angle regression. *Annals of Statistics*, 32, 407499.

Ernst J, Vainas O, Harbison CT, Simon Itamar and Bar-Joseph Z (2007): Reconstructing dynamic regulatory maps. *Molecular Systems Biology*, 3 Article number: 74.

Fan J and Li R (2001): Variable slection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, 96: 1348-1360.

Friedman J (2001): Multivariate adaptive regression splines. *Annals of Statistics*, 19:1-141.

Friedman JH and Popescu BE (2004): Gradient directed regularization. *Technical report, Stanford University.*

Friedman N (2004): Inferring cellular networks using probabilistic graphical models. *Science*, 30:799-805.

Gao F, Foat BC and Bussemaker HJ (2004): Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, 5, 31.

Gardner TS, di Bernardo D, Lorenz D and Collins J (2003): Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301: 102-105.

Gui J and Li J (2005): Threshold gradient descent method for censored data regression, with applications in pharmacogenomics. *Pacific Symposium on Biocomputing*, 10: 272-283.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE and Ren B (2007): Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39: 311-318.

Hong F and Li H (2006): Functional hierarchical models for identifying genes with different time-course expression profiles. *Biometrics*, 62: 534-544.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001): Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292: 929-34.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001): Lethality and centrality in protein networks. *Nature*, 411: 41-2.

Kanehisa M and Goto S (2002): KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28: 27-30.

Keles S, Van Der Laan M and Eisen MB (2002): Identification of regulatory elements using a feature selection method. *Bioinformatics* 18, 1167-1175.

Lee S, Pe'er D, Dudley AM, Church GM and Koller D (2007): Identifying regulatory mechnisms using individual variation reveals key role for chromatin modification. *Proceedings of National Academy of Sciences* U S A, 103: 14062-14067.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al.* (2002): Transcriptional regulatory networks in *S. cerevisiae*. *Science*, 298: 799-804.

Li H and Gui Li (2006): Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7: 302-317.

Luan Y, Li H (2004): Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics*, 20:332-339.

Ma P, Castillo-Davis C, Zhong W, and Liu JS (2006): A data-driven clustering method for time course gene expression data. *Nucleic Acids Research*, 34(4), 1261-1269.

Meinshausen N and Buhlmann P (2006): Consistent neighbourhood selection for high-dimensional graphs with the Lasso. *Annals of Statistics*, 34: 1436-1462.

McCullagh P and Nelder J (1989): *Generalized Linear Models*, 2nd ed., Chapman & Hall.

McLachlan GJ and Basford KE (1988): Mixture Models: Inference and Applications to Clusering. Marcel Dekker.

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS and Cheung VG (2004): Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430: 743747.

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M and Zhang C, et al. (2005): An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* 37: 710717.

Schafer J and Strimmer K (2005): An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21: 754-764.

Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field A, Moore IK, Wang JZ, Widom J (2006): A Genomic Code for Nucleosome Positioning. *Nature*, 442(7104):772-8

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003): Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2): 166-76.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D and Futcher B (1998): Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of Cell*, 9, 3273-3297.

Steifeld I, Shamir R and Kupiec M (2007): A genome-wide analysis in Saccharomyces cerevisiae demonstrates the influence of chromatin modifiers on transcription. *Nature Genetics*, 9(3):303-9

Storey JD, Xiao W, Leek JT, Dai JY, Tompkins RG and Davis RW (2005): Significance analysis of time course microarray experiments. *Proceedings of National Academy of Sciences*, 102, 12837-12842.

Tai YC and Speed TP (2006): A multivariate empirical Bayes statistic for replicated microarray time course data. *Annals of Statistics*, 34: 2387-2412.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ and Church GM (1999) Systematic determination of genetic network architecture. *Nature Genetics*, 22: 281-285.

Tegner J, Yeung MK, Hasty J, Collins JJ (2003): Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proceedings of National Academy of Sciences* U S A, 100: 5944-9.

Tibshirani R (1996): Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society (B)*, 58: 67-288.

Wang L, Chen G and Li H (2007): Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, in press.

Wei Z and Li H (2007a): Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics*, in press.

Wei Z and Li H (2007b): A Markov random field model for network-based analysis of genomic data. submitted.

Wille A, Zimmermann P, Vranova E, Furholz A, Laule, O, Bleuler, S, Hennig L, Prelic A, von Rohr P, Thiele L, Zitzler E, Gruissem W. and Bhlmann P (2004): Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biology*, 5(11) R92, 1-13.

Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, and Rando OJ (2005): Genome-scale Identification of Nucleosome Positions in S. Cerevisiae. *Science*, 309(5734), 626-630.

Yuan M and Kendziorski C (2006): Hidden Markov models for microarray time course data in multiple biological conditions. *Journal of American Statistical Association*, 101(476), 1323-1340.

Yuan M and Lin Y (2006): Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society B*, 68: 49-67.

Zou H and Hastie T (2005): Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society SER B-STAT MET*, 67:301-320.

Zou H, Hastie T and Tibshirani R (2007): On the "Degrees of Freedom" of the Lasso. *Annals of Statistics*, in press.